Handling Missing Strain (Rate) Curves Using K-Nearest Neighbor Imputation

Mahdi Tabassian*§, Martino Alessandrini*§, Ruta Jasaityte*, Luca De Marchi§, Guido Masetti§, and Jan D'hooge*
* Lab on Cardiovascular Imaging and Dynamics, Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium
§ Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, Italy
mahdi.tabassian@kuleuven.be

***Abstract*—Although a lot effort has been devoted over the past years to the accurate measurement of echocardiographic deformation curves in order to quantify regional myocardial function, much less attention has been paid to the problem of dealing with missing or artifactual curves. Considering the difficulties associated with missing or unreliable curves in the clinical diagnostic process, this study sought to examine the usefulness of the K-nearest neighbor (KNN) imputation algorithm to address this problem. Experiments with segmental strain (rate) curves of 30 normal subjects showed that the imputation algorithm can lead to low estimation errors even with a high percentage of missing data.**

I. Introduction

Echocardiographic strain (rate) imaging is the modality of choice for the noninvasive assessment of regional myocardial function. Quantification of regional heart function is performed by dividing the left ventricular (LV) walls into segments and measuring myocardial deformation in each of them by tissue Doppler or speckle tracking imaging techniques [1].

Like any imaging modality, echocardiographic deformation imaging has its own limitations and technical challenges. One of these challenges is that for having reliable deformation curves, the acquired images should be of good quality. Poor image quality- which can be due to artifacts, noise, aliasing, etc. - may lead to strain (rate) curves which are artifactual and even not physiologically meaningful. Such traces are not of use and cannot be analyzed by a cardiologist or an automatic classification algorithm. Moreover, acquiring images of all myocardial walls is not always possible. This can happen because of shadowing by ribs or lung tissue. As a consequence, strain (rate) curves are missing for some of the LV segments.

The aforementioned issues can reduce the diagnostic value of strain (rate) imaging as both the cardiologist or a computer-aided diagnostic system miss crucial information. An example for the latter is a machine learning algorithm that we have recently proposed to aid clinical decision-making [2]. This algorithm works based on extracting spatio-temporal characteristics of echocardiographic strain (rate) traces derived from 18 LV segments [3] in order to classify patients to healthy or pathological. Therefore, absence of any LV segment, and its corresponding parameters, prevents our method to be used in the clinical diagnostic process.

Substitution of missing values by the average of their corresponding available values or simply by zeros are two common approaches [4] that can be used to tackle the missing curve problem. Although these strategies are easy to implement, they do not lead to optimal estimations of missing values because relationships between data attributes are not taken into account.

An alternative is to use an *imputation* technique [4] which estimates the missing values of a given data vector by considering the similarities between its available values and those of a data set with complete attributes. The aim of the current study was therefore to evaluate, for the first time, the efficacy of a data imputation method to cope with the missing deformation curves problem.

The rest of this paper is organized as follows. Section II describes the utilized imputation technique as well as the employed evaluation metric and presents the details of data preparation and parameter setting. Results are given in Section III and Section IV discusses the results. Finally, Section V draws conclusions and summarizes the paper.

II. Materials and Methods

A. K-Nearest Neighbor Imputation

We used the K-nearest neighbor (KNN) imputation approach [5] in our experiments. This imputation technique has been proven to be an effective approach to estimate missing attributes of data sets with different characteristics and to be robust against high percentages of missing values [5]–[7].

Fig. 1 shows the concept of missing curves estimation using the KNN imputation approach where the absent measurements of a LV wall are filled in with the outcome of the imputation phase. The KNN imputation technique uses a group of instances without any missing data (e.g. a group of subjects with deformations curves of all 18 LV segments) to estimate missing values of a given instance $\boldsymbol{x}$. The imputation algorithm works by searching for K most similar instances of the complete data group to the given sample $\boldsymbol{x}$ by discarding those variables of the complete data which are missing in $\boldsymbol{x}$. The instances in the KNN set are then used to compute a weighted average of their values that are missing in $\boldsymbol{x}$ and the result is considered as an estimate of the absent measurements.

To apply the imputation algorithm to the missing segmental curves problem, we built two separate matrices of complete data, one for the strain traces and one for the strain rate traces, by concatenating all 18 available curves of each subject and

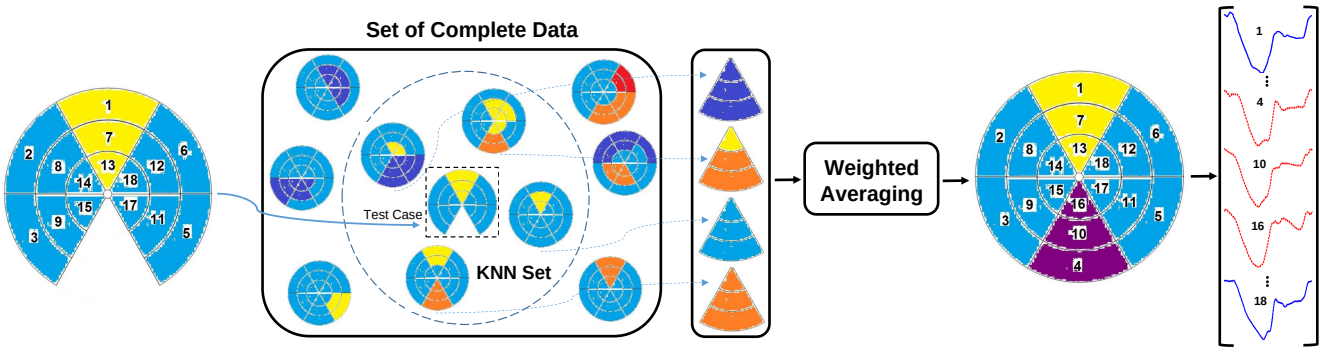

Fig. 1. Schematic illustration of the process of estimating missing segmental curves of a LV wall by the KNN imputation algorithm. After measuring the similarities between the available curves and the corresponding traces of a complete data set, a weighted average of a set of K most similar subjects' traces (red dashed curves) is computed to fill in the missing measurements.

superimposing the obtained vectors of all subjects. For a given subject $\boldsymbol{x}$ with some missing curves, its available curves were concatenated and the same was done for the corresponding curves of each vector of the complete matrix. Similarities between the obtained vector for $\boldsymbol{x}$ and those of the complete matrix were then measured using the Euclidean distance metric. Finding the K most similar vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$, a weight value for each of them was computed as follow:

$$w_i = \frac{\frac{1}{(d_i + \epsilon)}}{\sum_{j=1}^{K} \frac{1}{(d_j + \epsilon)}} \tag{1}$$

where d_i is the distance between $\boldsymbol{x}$ and the ith sample in the KNN set and ϵ is a small constant which ensures that an infinite weight value is not assigned to $\boldsymbol{x}$ if $d_i = 0$. Since w_i is proportional to the inverse of d_i, samples of the KNN set which are closer to $\boldsymbol{x}$ have greater contributions in approximating the missing values.

B. Data Acquisition and Preprocessing

A set of 30 normal subjects from the DOPPLER-CIP study [8] was used in our experiments. Color tissue Doppler data acquisition was performed at high frame rate (>180 Hz) with a GE VingMed Vivid E9. For each subject, data were acquired in the apical 2-, 3- and 4-chamber views and Doppler-based longitudinal strain (rate) traces were extracted as previously described in [9]. Since the number of samples of the extracted curves could be different due to the differences in the subjects' heart rates, a linear interpolation procedure was adopted to have the same number of samples in all traces. Each of the six mechanical phases of the cardiac cycle (i.e. electromechanical coupling, isovolumetric contraction, ejection, isovolumetric relaxation, early filling and late filling [10]) was interpolated separately and then merged to have the whole heart cycle.

C. Generating Data with Missing Values

First, the data was divided into a training and a testing set. Then, the matrix of complete data was built with the subjects of the training set and the testing set was used to create cases with missing data. In order to generate data with different percentages of missing values, the segmental strain (rate) curves of each subject were removed in three different ways:

1) one out of 18 segments was deleted (5.55% data missing), representing an acquisition of an artifactual segment,
2) one out of six LV walls (i.e. three segments) was deleted (16.66% data missing), representing a shadowed wall,
3) two contralateral walls were deleted (33.33% data missing), representing a missing apical view.

For each of the above cases, all possible selections were examined and the removed segmental curves were stored for comparison with the imputed ones.

D. Parameter Setting with Cross-Validation

The cross-validation (CV) technique was adopted to find the optimal number of KNNs in the imputation phase. Experiments were carried out with 10-fold CV by randomly dividing the subjects into 10 equal-size folds. In each round of CV, nine folds were used as the training data and the last one was used as the testing set. This procedure was repeated 10 times so that all folds were used in the training and test sets. Average outcome obtained in the 10 rounds of CV was considered to assess the imputation algorithm's performance.

E. Evaluation of the Imputation Algorithm

Performance of the KNN imputation algorithm was evaluated by comparing the original and imputed segmental curves for each value of K and computing a relative-root-sum-of-squares-error (RRSSE) value as follow:

$$RRSSE = \frac{\sqrt{\sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2}}{A_{|\boldsymbol{y}|}} \tag{2}$$

where $\boldsymbol{y}$ is the original curve with n samples, $\hat{\boldsymbol{y}}$ is the imputed curve and $A_{|\boldsymbol{y}|}$ denotes the area under the absolute curve of the original data. The rationale for using $A_{|\boldsymbol{y}|}$ in (2) was to take into account the amplitudes of the original curves throughout

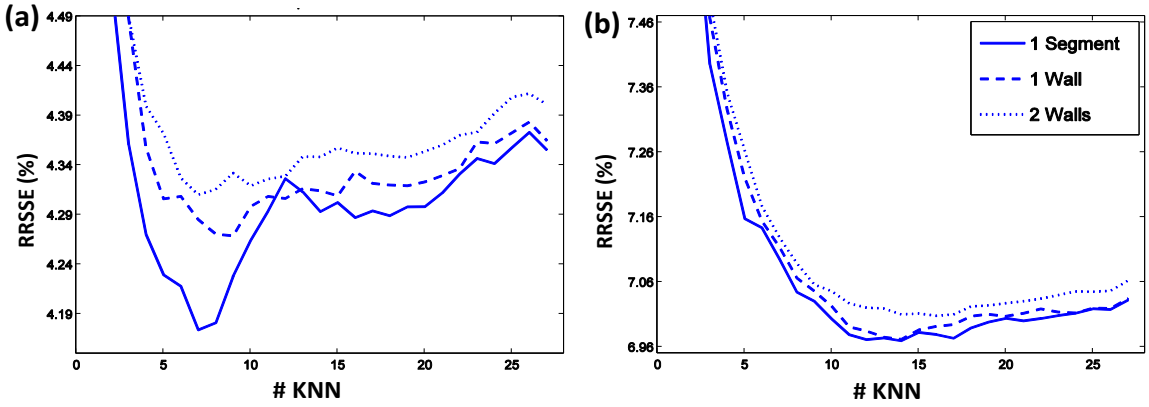

Fig. 2. The average RRSSE values (%) of 10-fold CV over different values of K for the three cases of curve removal. (a) Estimation errors for the missing strain and (b) strain rate curves.

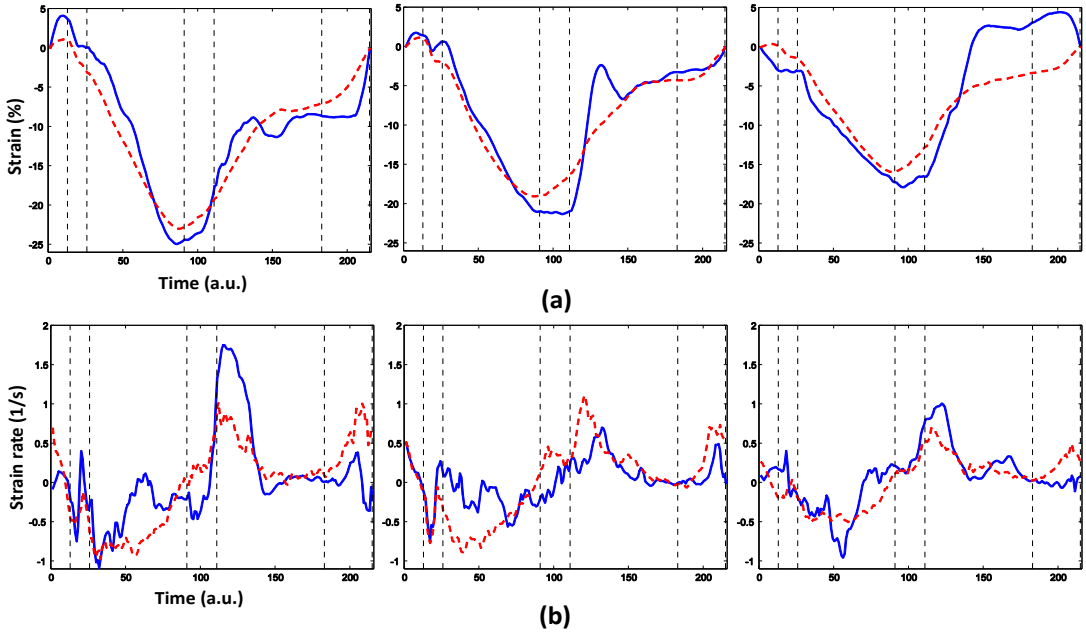

Fig. 3. An example of original (solid blue) and imputed (dashed red) segmental (a) strain and (b) strain rate curves of one of the LV walls obtained with optimal K values. The six mechanical phases of the cardiac cycle are shown with dashed vertical lines in order to facilitate comparison of the original and imputed curves' patterns.

the cardiac cycle. Assuming that sums of squared differences between two imputed segmental curves $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ and their original traces $\boldsymbol{y}$ and $\boldsymbol{z}$ are the same, a smaller RRSSE value is assigned to the imputed curve in which its original trace has a larger absolute amplitude throughout the cardiac cycle.

The examined values of K in the imputation phase were taken from the interval $[1, 27]$ where the maximum value was the number of the training subjects.

III. Results

Fig. 2 demonstrates average RRSSE values of 10-fold CV for the imputed strain and strain rate curves and for the three cases of curve removal, mentioned in Section II-C, over different values of K. It shows that the imputation algorithm yielded its best performance with K within the range $5-8$ for the strain curves and $10-15$ for the strain rate traces. It is also evident from the figure that by increasing the number of the missing curves, imputation errors increased, although the increments were more pronounced for the strain curves.

The lowest average estimation errors for the missing strain and strain rate curves are listed in Table I. In order to statistically compare these estimation errors, a paired t-test was done and the results are also shown in Table I. It can be seen that the estimation errors for the different cases of the missing curves are not statistically significant for p-value < 0.05.

In order to give further insight into the performance of the imputation algorithm, an example of the original and

TABLE I

THE LOWEST AVERAGE ESTIMATION ERRORS AND THEIR STANDARD DEVIATIONS (%) FOR THE IMPUTED STRAIN AND STRAIN RATE CURVES AND THE THREE CASES OF EXAMINED CURVE REMOVALS. THE ESTIMATION ERROR VALUES OF THESE THREE REMOVAL CASES WERE COMPARED WITH THE PAIRED t-TEST METHOD TO SEE WHETHER THEIR DIFFERENCES ARE STATISTICALLY SIGNIFICANT.

	Mean ± STD	1 Seg. vs. 1 Wall	1 Seg. vs. 2 Walls	1 Wall vs. 2 Walls
Strain				
1 Seg.	4.17 ± 4.45			
1 Wall	4.27 ± 3.10	p = 0.79	p = 0.77	p = 0.91
2 Walls	4.31 ± 2.11			
Strain Rate				
1 Seg.	6.97 ± 2.27			
1 Wall	6.97 ± 1.55	p = 0.99	p = 0.87	p = 0.84
2 Walls	7 ± 1.16			

imputed strain and strain rate curves for one of the LV walls is shown in Fig. 3. One can see that the imputed curves, which obtained using the best K values shown in Fig. 3, follow the patterns of their corresponding original curves. More specifically, decreasing trend in the maximum absolute amplitudes of the original strain curves (Fig. 3 (a)) can also be observed for the imputed traces and timings of the important events of the strain rate profiles (e.g. peak-early diastolic strain rate and peak strain rate during atrial filling) are comparable for the original and imputed curves (Fig. 3 (b)).

IV. DISCUSSION

This study aimed to examine the usefulness of the KNN imputation approach for dealing with the problem of missing segmental strain (rate) curves. All possible cases of missing only one segmental curve and curves of one wall or two contralateral walls, as the most frequent types of missing curves in echocardiography, were investigated in our experiments.

The small estimation error values of $\sim 4.2 - 4.3\%$ for the strain curves and $\sim 7\%$ for the strain rate traces (c.f. Table I) demonstrate the capability of KNN imputation in accurately finding subjects with deformation curves similar to those of a given subject with missing curve(s). Given that the best performances of the imputation algorithm for the strain and strain rate traces were achieved within a range of nearest neighbors, finding the optimal value of K is not a main challenge in the imputation process.

The imputation algorithm also appeared to be robust against increasing the number of missing curves where the estimation errors for all three removal cases were comparable and not significantly different. Beside the robustness of the KNN imputation method, which has already been shown in the literature, diversity in temporal behavior of the different LV segments' deformation curves could also aid the imputation algorithm to correctly find the most identical subjects. Having diverse segmental curves implies that by increasing the number of missing traces, the present curves may still represent distinct patterns. This can increase the chance of finding those subjects which were existing in the KNN set when a smaller number of curves were missing.

The fact that the imputed curves, shown in Fig. 3, mimic the patterns of the original traces suggests that the widely used conventional markers (like peak-systolic strain, end-systolic strain, peak-systolic strain rate and peak-early diastolic strain rate values) can be extracted from the estimated curves to give clinicians an idea of the values that these markers could have in case their corresponding curves were not missing. Utility of the imputed markers and a careful comparison with the original measurements, however, remain to be validated by further prospective studies. In our future research, we will also apply the KNN imputation method to a database containing pathological cases in order to investigate its capability in estimating missing abnormal deformation curves.

V. CONCLUSIONS

We have addressed, to the best of our knowledge for the first time, the issue of missing echocardiographic deformation curves using the KNN imputation algorithm. Quantitative and qualitative analyses of the experimental results showed that this imputation technique is capable of handling the problem of missing strain (rate) curves. In particulate, the imputation errors were small, their differences for the different percentages of the missing curves were not statistically significant and the imputed curves could mimic the patterns of the original traces.

REFERENCES

[1] V. Mor-Avi, R. M. Lang, L. P. Badano *et al.*, "Current and evolving echocardiographic techniques for the quantitative evaluation of cardiac mechanics: ASE/EAE consensus statement on methodology and indications: endorsed by the Japanese society of echocardiography," *Journal of the American Society of Echocardiography*, vol. 24, no. 3, pp. 277–313, 2011.

[2] M. Tabassian, M. Alessandrini, L. Herbots *et al.*, "Automatic detection of ischemic myocardium by spatio-temporal analysis of echocardiographic strain and strain rate curves," in *IEEE International Ultrasonics Symposium (IUS)*, 2015, pp. 1–4.

[3] R. M. Lang, L. P. Badano, V. Mor-Avi *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging," *European Heart Journal - Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[4] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," *NY Springer*, 2001.

[5] O. Troyanskaya, M. Cantor, G. Sherlock *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[6] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, clustering, and data mining applications*. Springer, 2004, pp. 639–647.

[7] S. G. Liao, Y. Lin, D. D. Kang *et al.*, "Missing value imputation in high-dimensional phenomic data: imputable or not, and how?" *BMC bioinformatics*, vol. 15, no. 1, pp. 1–12, 2014.

[8] F. Rademakers, J. Engvall, T. Edvardsen *et al.*, "Determining optimal noninvasive parameters for the prediction of left ventricular remodeling in chronic ischemic patients," *Scandinavian Cardiovascular Journal*, vol. 47, no. 6, pp. 329–334, 2013.

[9] P. Claus, J. D'hooge, T. Langeland *et al.*, "SPEQLE (Software Package for Echocardiographic Quantification LEuven) an integrated approach to ultrasound-based cardiac deformation quantification," in *IEEE Comput. Cardiol.*, 2002, pp. 69–72.

[10] J. D'hooge, B. Bijnens, J. Thoen *et al.*, "Echocardiographic strain and strain-rate imaging: a new tool to study regional myocardial function," *IEEE Trans. Med. Imaging*, vol. 21, no. 9, pp. 1022–1030, 2002.